USE OF LINEAR EXTRAPOLATION BASED LINEAR PREDICTIVE CEPSTRAL FEATURES (LE-LPCC) FOR TAMIL SPEECH RECOGNITION

Rathinavelu Chenganavarayan
Speech Processing Group,
Lucent Speech Solutions Department
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

A new method, named linear prediction with linear extrapolation has been proposed in the past, which aims at modifying conventional linear prediction especially for speech coding applications. The basic idea is to reformulate the computation of linear prediction so that an optimal FIR-predictor of order 2p could be determined from p numerical values. In this work, we extend the above method to generate the cepstral features using the compressed LPC parameters and use the new feature for accurate speech recognition. Preliminary results on Tamil connected digits recognition task demonstrated that the cepstral features derived from the new approach yield more accurate modeling of speech spectra and provides better discrimination among different speech classes.

1. INTRODUCTION

Feature extraction is the process of deriving a compressed set of features that are characteristic of a given signal. These features are desired to preserve all the information relevant to the application, and to have no redundancy in representing the signal. One of the most powerful speech analysis techniques is the method of linear predictive (LPC) analysis [6]. The LPC analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint. It is faster and provides extremely accurate estimates of speech parameters if the signal can be well modeled by the all-pole model [14]. Linear prediction parameters have been used useful in a variety of applications, such as speech coding, speech recognition, speech synthesis, speech enhancement and speaker recognition [4, 12, 5, 3, 8].

Tamil is the oldest attested member of the Dravidian family of languages, which is predominant in southern India. Tamil Nadu, the southernmost state of the Indian subcontinent, is the cradle of the oldest and largest Indian literature and culture. Apart from being the language of fifty million people in Tamil Nadu, Tamil is the spoken and written languages of several millions of Tamils living in Sri Lanka, Indonesia, Singapore, Malaysia, Burma, Mauritius, Fiji Islands, South Africa, Australia, Canada, UK, USA, New Zealand, and various countries in Europe [1]. The alphabet of Tamil is unique, and is like English in that it is phonetic. That is, letters represent sounds, rather than ideas as in Mandarin Chinese where the intonation of a word determines its meaning [16]. Language identification using Tamil language has been studied recently by several researchers but not in speech recognition [13, 11]. This paper deals with a new set of LPC-derived features that is applied to Tamil connected digits speech recognition.

Recently, a new method, named linear prediction with linear extrapolation (LPLE), has been proposed which aims at modifying conventional linear prediction especially for speech coding applications [2]. The idea is to reformulate the computation of linear prediction so that an optimal FIR-predictor of order 2p could be determined from p numerical values. In this study we explore the use of newly derived linear extrapolation based linear predictive cepstral coefficients (LE-LPCC) as speech features in a speaker-independent recognition experiment. Preliminary results demonstrated that the cepstral features derived from the new approach yield more accurate modeling of speech spectra and provides better discrimination among different speech classes in comparison to conventional linear predictive cepstral coefficients.

2. LPLE-DERIVED CEPSTRUMS

Let \( x_n \) be a sample to be predicted from its \( p \) previous values. In conventional linear prediction \( p \) preceding samples of \( x_n \) are taken as such for linear combination. In the case of LPLE method, the \( x_n \) is predicted by using \( p \) number of sample pairs that occur before time instant \( n \). Each sample pair is connected by a line. These lines are determined by arranging \( 2p \) preceding values of \( x_n \) into \( p \) groups of two consecutive samples. The equation for the line that connects two consecutive samples \( x_{n-2p} \) and \( x_{n-2p+1} \) can be expressed as follows:

\[
J_n = 2i \left[ x_{n-2p+1} - x_{n-2p} \right] + x_{n-2p} \quad (1)
\]

Linear extrapolation, each of the lines that are determined by two consecutive samples, is then used to compute the value of the line at time instant \( n \) as shown in Figure 1. By combining all the sample pairs \( p \) extrapolated values are obtained at time instant \( n \). A linear combination of

\[
J_n = 2i \left[ x_{n-2p+1} - x_{n-2p} \right] + x_{n-2p} \quad (1)
\]

Linear extrapolation, each of the lines that are determined by two consecutive samples, is then used to compute the value of the line at time instant \( n \) as shown in Figure 1. By combining all the sample pairs \( p \) extrapolated values are obtained at time instant \( n \). A linear combination of
Figure 1. Calculation of speech samples for prediction of x(n).

The predicted value of the n-th speech sample is given by

$$x_n = \sum_{i=1}^{p} a_i (x_{n-2i+1} - x_{n-2i}) + x_{n-2i}$$

(2)

where $a_i$ are the predictor coefficients and $x_n$ is the n-th speech sample. The value of $p$ required for adequate modeling of the vocal tract depends on the sampling frequency used in digitization of the signal: the higher the sampling frequency, the larger the analysis order $p$ should be. It has been suggested that when the sampling frequency in kHz is $q$ then the analysis order should be at least $q+1$ in case of conventional linear prediction.

The prediction error, the difference between the predicted value and the actual measured value, is defined by

$$e_n = x_n - \sum_{i=1}^{p} a_i (x_{n-2i+1} - x_{n-2i}) - x_{n-2i}$$

(3)

The total squared error can be denoted by $E$, where

$$E = \sum_{n} c_n$$

(4)

The LPC coefficients which minimize the mean squared prediction error can be obtained by setting

$$\frac{\partial E}{\partial a_i} = 0, \quad 1 \leq i \leq p.$$

(5)

The following $p$ normal equations can be solved by using Gaussian elimination schemes [15]:

$$\sum_{i=1}^{p} a_i \left[ (-4i + 2j) r_{2i-2j-1} + (8ij - 2i - 2j + 1) r_{2i-2j} \right]$$

$$+ \left[ (-4i + 2j) r_{2i-2j+1} \right]$$

$$= -2j r_{2i-1} + (2j - 1) r_{2i}, \quad 1 \leq i \leq p$$

where the autocorrelation function of $x_n$, denoted by $r_n$

$$r_i = \sum_{n=0}^{N-1-i} x_n^w x_{n+i}^w$$

(6)

and $x_n^w = x_n \cdot w_n$, and $w_n$ is the hamming window weighting function:

$$w_n = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right), \quad 0 \leq n \leq N - 1.$$

(7)

The transfer function of the all-pole filter given by LPLE-analysis can now be expressed as:

$$H_z = \frac{1}{1 + \sum_{i=1}^{p} [2i a_i z^{-2i+1} + (1 - 2i) a_i z^{-2i}]}$$

(8)

It can be seen from the above equation that the LPLE method yields an all-pole filter of the order $2p$ when the number of the normal equations is equal to $p$. Since the denominator of $H_z$ has $2p$ complex roots, the spectra of $H_z$ has $p$ resonant frequencies, which correspond to formant frequencies. This implies that the LPLE technique can model the spectrum of the vocal tract as a spectrum of an order-$2p$ all-pole model $H_z$. Cepstral coefficients, which can be obtained from conventional LPC analysis, have been widely used in speech recognition. The cepstral coefficients, $c_n$, of the spectra obtained from LPLE analysis can also be computed recursively from the LPC coefficients as follows.

If all the poles of $H_z$ are inside the unit circle, in $H_z$ can be expressed as

$$\ln H_z = \sum_{n=1}^{\infty} c_n z^{-n}$$

(9)

To obtain a simple and unique relationship between the parameters $c_n$ and $a_n$ by substituting $H_z$ into the above equation and by taking derivatives on both sides with respect to $z^{-1}$:

$$\frac{d}{dz^{-1}} \ln \left( 1 + \sum_{i=1}^{p} [2i a_i z^{-2i+1} + (1 - 2i) a_i z^{-2i}] \right)$$

$$= \frac{d}{dz^{-1}} \left( \sum_{n=1}^{\infty} c_n z^{-n} \right)$$

(10)

which is simplified to

$$-\sum_{n=1}^{p} \left[ 2i a_i z^{-2i-1} + (1 - 2i) a_i z^{-2i} \right]$$

$$= \sum_{n=1}^{\infty} n c_n z^{-n+1}$$

(11)

and rewritten as

$$-\sum_{n=1}^{\infty} \left[ 2i a_i z^{-2i-1} + (1 - 2i) a_i z^{-2i} \right]$$

$$= \sum_{n=1}^{\infty} n c_n z^{-n+1}$$

(12)
Figure 2. Block diagram for the calculation of the LE-LPCC cepstrum.

If we equate the constant term and the various powers of $z^{-1}$ on the left and right sides of the above equation, we obtain the desired relationship between $c_n$ and $a_n$. Even though there is an infinite number of cepstrum coefficients, we truncate the sequence to 12. This indicates that the characteristics of vocal tract and excitation are well represented separately in the cepstral domain. The higher order coefficients take the excitation property and the lower order coefficients take the vocal tract property. For simplicity, we assume the LPC order $p$ to be 3, then the equation (12) becomes

$$-2a_1 + 2a_1 z^{-1} - 12a_2 z^{-2} + 12a_2 z^{-3} - 30a_3 z^{-4} + 30a_3 z^{-5} = (c_1 + 2c_2 z^{-1} + 3c_3 z^{-2}) \left( 1 + 2a_1 z^{-1} - a_1 z^{-2} + 4a_2 z^{-3} - 3a_3 z^{-4} + 6a_3 z^{-5} - 5a_3 z^{-6} \right)$$

And the corresponding cepstral coefficients are obtained by equating the powers of $z$ on both sides of the above equation:

$$c_1 = -2a_1$$
$$c_2 = a_1 - a_1c_1$$
$$c_3 = \frac{-2a_2 + a_1c_1 - 4c_2a_1}{3}$$

Similarly the higher order cepstral coefficients are computed in a recursive way. A straightforward algorithm is shown in Figure 2 to calculate the new feature set from the given speech signal. Equation (12) allows us to compute the coefficients $c_n$ from the $p$ predictor coefficients and the predictor coefficients from the $p$ coefficients $c_1, c_2, \ldots, c_p$.

3. EXPERIMENTAL SETUP

Eleven Tamil digits, onnu, erandhu, mooru, maangi, ainthu, aaru, ezhu, ettu, onpathu and poochiam as shown in Table 1 were used in the evaluation task. The Tamil connected digit corpus comprised of speech data from 16 native speakers (8 male and 8 female ranging in age from 15 to 75 years old) collected over network channels using a variety of telephone handsets. The digit string length is fixed to 10 for both training and testing sets. Each talker spoke a total of 150 digit strings, out of which the first 100 utterances were used for training and the remaining 50 utterances were used for testing. All recordings in the training and testing set are valid digit strings, totaling 1000 and 800 strings for training and testing, respectively.

Input speech is segmented into overlapping frames of 30 msec long with centers 10 msec apart. The basic recognizer feature set, denoted as, LPCC, consists of 39 features that includes the 12 liltered conventional linear predictive cepstral coefficients, the normalized log-energy, the first and second order derivatives of cepstrums and log-energies [8]. LE-LPCC contains the same dimension as LPCC feature set but the cepstrums are generated from linear extrapolation based linear predictive coefficients as explained in the previous section. Make a note that we used an LPC order of eight for both analysis methods (i.e., $p = 8$).

### Table 1. One-to-one mapping between English and Tamil digits.

<table>
<thead>
<tr>
<th>Digit</th>
<th>English</th>
<th>Tamil</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>one</td>
<td>onnu</td>
</tr>
<tr>
<td>2</td>
<td>two</td>
<td>erandhu</td>
</tr>
<tr>
<td>3</td>
<td>three</td>
<td>mooru</td>
</tr>
<tr>
<td>4</td>
<td>four</td>
<td>maangi</td>
</tr>
<tr>
<td>5</td>
<td>five</td>
<td>ainthu</td>
</tr>
<tr>
<td>6</td>
<td>six</td>
<td>auru</td>
</tr>
<tr>
<td>7</td>
<td>seven</td>
<td>ezhu</td>
</tr>
<tr>
<td>8</td>
<td>eight</td>
<td>ettu</td>
</tr>
<tr>
<td>9</td>
<td>nine</td>
<td>onpathu</td>
</tr>
<tr>
<td>0</td>
<td>zero</td>
<td>poochiam</td>
</tr>
<tr>
<td>O</td>
<td>oh</td>
<td>suzhi</td>
</tr>
</tbody>
</table>

4. EXPERIMENTAL RESULTS

Following feature analysis, each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-independent models. Each digit was modeled with 10 state HMMs, with 4 Gaussian mixtures. Speech background is modeled with a single state, 32 mixture HMM. The HMMs were trained using one iterations of maximum-likelihood estimation (MLE) and six iterations of minimum string error (MSE) training [9]. Each training utterance is signal conditioned by applying cepstral mean subtraction prior to being used in both training methods [10].

We have conducted experiments to verify the effectiveness of the proposed LE-LPCC derived feature extraction tech-
Table 2. Word error rate (WdEr) and string accuracy (StAc) for a 10-digit known-length grammar-based Tamil connected digit recognition task using the conventional MLE and MSE training methods as a function of feature type.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>MLE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (LPCC)</td>
<td>8.49%</td>
<td>67.71%</td>
</tr>
<tr>
<td>Proposed (LE-LPCC)</td>
<td>7.47%</td>
<td>72.79%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

A new feature, named linear prediction with linear extrapolation based cepstral coefficients, is proposed. We also described the detailed mathematical derivation of cepstral generation process from the compressed LPC parameters and provided the implementation issues of the frontend feature extraction process. Experimental results on Tamil connected digit recognition task demonstrated that the cepstral features derived from the new approach yield more accurate modeling of speech spectra and provides better discrimination among different speech classes. In comparison to the conventional LPCC, the proposed LE-LPCC features are able to yield more accurate models especially for the formant structure of speech in the case when the number of unknowns in the normal equations is small. However, when the number of parameters to determine the all-pole filter is larger, then the performance differences between the two feature models become small.

Acknowledgements

The author would like to thank all sixteen native Tamil speakers of Kadamattthur, Kanchipuram and Madras, India for providing their valuable speech recordings.

REFERENCES


